Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ:
заведующий кафедрой

_____ В. А. Кратасюк
“___” _____ 2020 г.

БАКАЛАВРСКАЯ РАБОТА

03.03.02 Физика

ГОМЕОБОКС-СОДЕРЖАЩИЕ ГЕНЫ В ИЗУЧЕНИИ ПРОЦЕССА РАЗВИТИЯ СОСНЫ ОБЫКНОВЕННОЙ *Pinus sylvestris*

Руководитель: _____ д.ф.-м.н., проф. М. Г. Садовский
дата, подпись; уч.степень, должность

Выпускник: _____ Т. А. Гусева
дата, подпись

Красноярск 2020

РЕФЕРАТ

Выпускная квалификационная работа содержит 28 страниц текстового документа, 64 использованных источников, 7 рисунков, 5 таблиц.

ГОМЕОБОКС-СОДЕРЖАЩИЕ ГЕНЫ, ДИФФЕРЕНЦИАЛЬНАЯ ЭКСПРЕССИЯ, СОСНА ОБЫКНОВЕННАЯ, ФУНКЦИИ ГЕНОВ, РАЗВИТИЕ ОРГАНИЗМА, ГОМЕОБОКС, ГОМЕОДОМЕН.

Цель работы — определение функций гомеобокс-содержащих генов в различных тканях *Pinus sylvestris*.

Предмет исследования — нуклеотидные последовательности, полученные в результате ассемблирования транскриптома *Pinus sylvestris* и обнаруженные в них гомеобокс-содержащие гены, их дифференциальная экспрессия в различных тканях.

Объект исследования — взаимосвязи между уровнями экспрессии гомеобокс-содержащих транскриптов и особенностями морфологического развития у разных видов хвойных.

Актуальность исследования обусловлена тем, что комплексный анализ функций и экспрессии гомеобокс-содержащих генов для вида *Pinus sylvestris* ранее не проводился.

Собран *de novo* транскриптом *Pinus sylvestris*, содержащий 775 502 транскрипта. Отобрано и аннотировано 243 гомеобокс-содержащих транскрипта. Определены дифференциально экспрессирующиеся гомеобокс-содержащие гены пяти различных тканей сосны обыкновенной и построена тепловая карта. Предположительно определены функции некоторых гомеобокс-содержащих генов: *HDG2*, *WOX3*, *WOX4*, *WOX8*, *WOX9*, *ATHB-13*, *HOX5*, *HOX20*, *HOX21*, *HAT7*, *PDF2*, *ROC1*, *ROC2*, *ROC8*, *BLH4*, *KN-1*, *KNAT2*, *KNAT1*, *KNAT3*.

СОДЕРЖАНИЕ

Введение **4**

1 Обзор литературы **7**

1.1 Актуальность исследования 7

1.2 Гомеобокс-содержащие гены 8

1.3 Гомеобокс-содержащие гены растений 9

1.4 Классификация гомеобокс-содержащих генов растений . . . 10

1.5 Изучение гомеобокс-содержащих генов растений 10

1.6 Биоинформатический анализ 12

1.6.1 Анализ данных РНК секвенирования 12

1.6.2 Данные РНК секвенирования 13

1.6.3 Сборка транскриптома *de novo* 13

1.6.4 Дифференциальная экспрессия генов 14

2 Материалы и методы **17**

2.1 Предобработка данных и *de novo* сборка транскриптома *P. sylvestris* 17

2.2 Оценка полноты сборки транскриптома 19

2.3 Отбор гомеобокс-содержащих транскриптов 19

2.4 Анализ дифференциальной экспрессии генов и построение тепловой карты 20

2.5 Аннотация дифференциально экспрессирующихся транскриптов 22

3 Результаты и обсуждение **23**

3.1 Сборка транскриптома и оценка качества 23

3.2 Классификация гомеобокс-содержащих транскриптов 23

3.3 Дифференциальная экспрессия гомеобокс-содержащих генов в различных тканях *Pinus sylvestris* 25

Список использованных источников **35**

ВВЕДЕНИЕ

Сосна обыкновенная *Pinus sylvestris* широко распространена по всей территории Российской Федерации и имеет важное экономическое и экологическое значение. *P. sylvestris* относится к группе голосеменных растений, семейство сосновые, рода сосна. На уровне геномики и транскриптомики данная группа изучена значительно меньше, чем группа покрытосеменных растений, поэтому изучение голосеменных представляет особый интерес.

Развитие *P. sylvestris* у других растений, принадлежащих семейству сосновых, изучается рядом ученых, однако на данный момент имеется сравнительно мало данных о генетических механизмах развития этого вида [1–3]. Комплексный анализ профиля дифференциальной экспрессии гомеобокс-содержащих генов является важным инструментом для функционального анализа и определения роли данных генов в процессе развития и роста растения.

Актуальность работы связана с тем, что подобный анализ данных РНК-секвенирования *P. sylvestris* ранее не проводился.

Объектом настоящей работы являются взаимосвязи между уровнями экспрессии гомеобокс-содержащих транскриптов и особенностями морфологического развития у разных видов хвойных.

Предметом данного исследования являются нуклеотидные последовательности, полученные в результате ассемблирования транскриптома *P. sylvestris* и обнаруженные в них гомеобокс-содержащие гены, их дифференциальная экспрессия в различных тканях.

Целью данной работы является определение функций гомеобокс-содержащих генов в различных тканях *Pinus sylvestris*. Иными словами, определить в каких процессах развития и дифференцировки тканей могут принимать участие гомеобокс-содержащие гены у сосны обыкновенной. Многие функции данных генов всё еще неизвестны (особенно у хвойных видов). Полученные данные помогут расширить и углубить понимание процессов развития и морфогенеза данного вида и семейства сосновых в целом. Для достижения поставленной цели был выполнен ряд **задач**, а именно:

1) Произведен отбор данных РНК-секвенирования различных тканей сосны обыкновенной;
2) Выполнены предобработка данных и оценка качества ридов;
3) Осуществлена *de novo* сборка транскриптома *P. sylvestris*;
4) Произведен отбор транскриптов, содержащих гомеобокс-содержащие гены;
5) Выполнен анализ дифференциальной экспрессии генов и их аннотация;
6) Определены функции гомеобокс-содержащих транскриптов.

Работа и связанные с ней исследования докладывалась на следующих конференциях:

1) 8$^{\text{th}}$ International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO), 2020 г, Гранада, устный доклад;
2) Международная конференция студентов, аспирантов и молодых ученых «Проспект Свободный — 2020», 2020 г, Красноярск, устный доклад, II место;
3) XI международная конференция «Dynamical Systems Applied to Biology and Natural Sciences» (DSABNS), 2020 г, Тренто, устный доклад;
4) Международная конференция летней школы по биоинформатике, 2020 г, Москва, Институт биоинформатики, устный доклад;
5) 7$^{\text{th}}$ International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO), 2019 г, Гранада, устный доклад;
6) Международная конференция студентов, аспирантов и молодых ученых «Проспект Свободный – 2019», 2019 г, Красноярск, устный доклад, III место;
7) X международная конференция «Dynamical Systems Applied to Biology and Natural Sciences» (DSABNS), 2019 г, Неаполь, устный доклад;
8) 56-я Международная научная студенческая конференция, 2018 г, Новосибирск, устный доклад.

Результаты работы и связанных с ней исследований опубликованы в следующих научных журналах и сборниках научных мероприятий:

- Гусева Т.А., Бирюков В.В. Гомеобокс-содержащие гены в изучении процесса развития сосны обыкновенной *Pinus sylvestris* // Проспект Свободный – 2020: материалы XVI Международной конференции студентов, аспирантов и молодых ученых. — 2020;
- Guseva T., Biriukov V., Sadovsky M. Role of Homeobox Genes in the Development of *Pinus sylvestris* // Lecture Notes in Computer Science, Springer Verlag. — 2020. — Vol. 12108 — Pp. 429 – 437;
- Guseva T., Sadovsky M., Biriukov V. Homeobox Genes: Investigating the Development of *Pinus sylvestris* (Scots pine) // 11^{th} International Conference Dynaamical Systems Applied to Biology and Natural Sciences: Book of Abstracts. — 2020. — Vol. 11. — Pp. 113-115.
- Садовский М.Г., Гусева Т.А., Бирюков В.В. Определение тканеспецифичности в тотальном транскриптоме лиственницы сибирской с помощью условной энтропии // Проспект Свободный – 2019: материалы XV Международной конференции студентов, аспирантов и молодых ученых, посвященной Международному году Периодической таблицы химических элементов Д. И. Менделеева. — 2019. — Стр. 682 – 684;
- Sadovsky M., Guseva T., Biriukov V. Entropy Approach to Identify Tissue Specificity in Total Transcriptome // 10^{th} International Conference Dynaamical Systems Applied to Biology and Natural Sciences: Book of Abstracts. — 2019. — Vol. 10. — Pp. 98-99;
- Sadovsky M., Guseva T., Biriukov V. Triplet Frequencies Implementation in Total Transcriptome Analysis// Lecture Notes in Computer Science, Springer Verlag. — 2019. — Vol. 11465 — Pp. 370 – 378;
- Гусева Т.А. Исследование особенностей структуры тотального транскриптома лиственницы сибирской на основе статистических свойств контигов // Материалы 56-й Международной научной студенческой конференции (МНСК). — 2018. — Т. 56. — Стр. 10.

1 Обзор литературы

1.1 Актуальность исследования

Гомеобокс-содержащие гены являются очень важными регуляторными генами в развитии и росте растений, животных, грибов [4]. Несмотря на сравнительно хорошую изученность данных генов, комплексный анализ функций и экспрессии гомеобокс-содержащих генов для вида *Pinus sylvestris* ранее не проводился. В настоящий момент имеются исследования для других видов, посвященные определенным классам гомеобокс-содержащих генов, к примеру класса HD-ZIP I, в связи с тем, что гены данного класса участвуют в реакциях организма на абиотический и биотический стресс [5]. Также изучались гены класса *WUSHEL* риса посевного. Исследование проводилось для различных тканей и органов на различных стадиях развития с целью выявления функций данных генов [6]. Имеются исследования, посвященные роли гомеобокс-содержащих генов для отдельных стадий развития в семействе хвойных [1, 7–9], однако мало изучена общая картина экспрессии всего множества классов этих генов в тканях почки, флоэмы и хвои сосны обыкновенной.

Подобный анализ может расширить и углубить понимание процессов, происходящих при развитии тканей сосны обыкновенной. Полученная информация может помочь исследователям влиять на характеристики данного вида путем регуляции экспрессии генов или каким-либо иным образом. Имеются патенты, связанные с улучшением свойств посевных культур посредством регуляции экспрессии гомеобокс-содержащих генов [10]. Определение функций гомеобокс-содержащих генов является важным шагом в исследовании развития тканей и органов организма и делает возможным изучение ряда других феноменов. Примером подобного явления может служить реакция растения на абиотический и биотический стресс [11].

1.2 Гомеобокс-содержащие гены

Регуляция транскрипции генов внутри клетки реализуется через специфическое взаимодействие транскрипционных факторов (ТФ) с генами-мишенями. ТФ обычно включают два домена. Первый домен — ДНК-связывающий, который распознает специфические последовательности в регуляторных областях генов-мишеней, а второй домен осуществляет межбелковые взаимодействия, например с другими транскрипционными факторами. ТФ могут активировать или подавлять экспрессию генов-мишеней, тем самым изменяя транскриптом клетки. Гомеобокс (ГБ) кодирует один из наиболее распространенных и наиболее описанных ДНК- связывающих доменов в эукариотических организмах — гомеодомен (ГД). Гомеобоксы были названы в честь гомеозизного эффекта, который вызывается мутацией или эктопической (нарушенной) экспрессией ГБ — это аномалия развития, при которой один сегмент тела развивается по подобию другого [12]. Гомеобокс-содержащие гены были впервые обнаружены у *Drosophila melanogaster* [13].

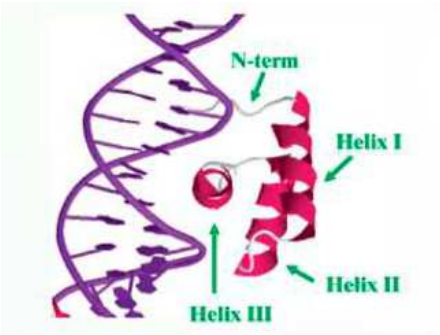

Рисунок 1 – Третичная структура гомеодомена [4]

Знание о механизмах специфичного связывания последовательности ДНК с ГД получено благодаря трехмерным структурам комплексов белок-ДНК в сочетании с направленным мутагенезом и биохимическим анализом [12, 14–16]. На данный момент достоверно известно, что ГД состоит из консервативного мотива длиной 60 аминокислот. Так как гомеодомен является весьма консервативным доменом, то по этой причине гомеобокс-содержащие гены могут быть обнаружены даже у немодельных организмов. Гомеодомен является частью транскрипционных факторов, в основном участвующих в процессах развития организма. ГД образует структуру из трех α-спиралей, I, II и III соответственно(Рисунок 2 и Рисунок 1). α-спирали соединены двумя «петлями» и одним «поворотом», участками, которые не являются спиральными. Гомеодомен способен связываться с ДНК с высоким сродством посредством взаимодействий, уста-

новленных спиралью III (называемой спиралью распознавания) с большой бороздкой ДНК-мишени. Одновременно с этим, неструктурированный N-концевой участок осуществляет взаимодействие с малой бороздкой ДНК (Рисунок 1) [15].

После открытия гомеодоменов в *Drosophila*, аналогичные ГД были обнаружены в транскрипционных факторах эволюционно далеких друг от друга групп организмов, таких как человек нематоды, грибы и растения. Высокая консервативность этого домена среди белков различных царств живых организмов говорит о том, что структура гомеодомена имеет важное значение для поддержания его функциональности, и указывает на существенную роль ГД в развитии организма [12, 17].

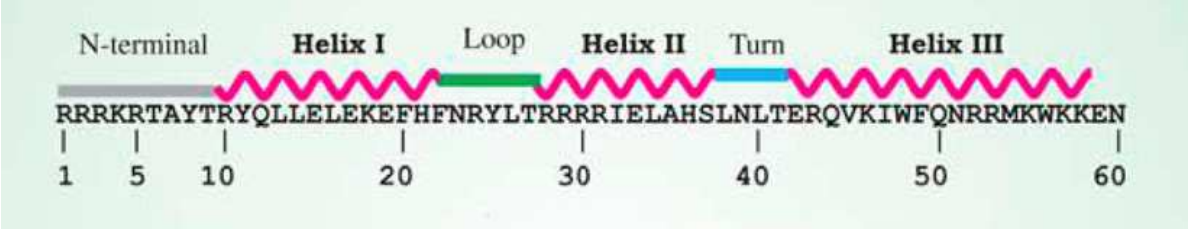

Рисунок 2 – Первичная и вторичная структура гомеодомена [4]

1.3 Гомеобокс-содержащие гены растений

ГБ-содержащие гены растений были впервые идентифицированы в лаборатории Эрика Фоллльбрехта в 1991 году [18]. Команда провела анализ мутантного образца кукурузы сахарной (*Zea mays* L.) с пораженными «узелковыми» листьями. Анализ такого мутанта привел к выделению *KNOTTED1*, первого ГБ-содержащего гена растений. С тех пор многие ГБ-содержащие гены были идентифицированы и выделены из множества растений, в том числе из классов однодольных и двудольных растений. Однако в отличие от ГБ-содержащих генов животных, ни один из известных растительных генов, содержащих ГБ, не обладает каноническим гомеозисным эффектом [19]. Канонический гомеозисный эффект определяется как превращение одной части организма в другую вследствие мутаций генов или их нарушенной экспрессии. Примером такого явления может служить

превращение антенн насекомых в ноги.

1.4 Классификация гомеобокс-содержащих генов растений

С момента открытия гена *KNOTTED1* у кукурузы, обнаружено большое количество генов, кодирующих ГД-содержащие ТФ, однако большинство из них до сих пор не описаны. ГД участвуют в самых разных процессах, например, в идентификации и поддержании меристематических клеток, эмбриогенезе, созревании пыльцы, регуляции развития цветков и реакции на условия окружающей среды [12]. Полное описание геномов некоторых растений, таких как, например, рис посевной (*Oryza sativa L.*), кукуруза сахарная (*Z. mays*), резуховидка Таля (*Arabidopsis thaliana L.*) и тополь волосистоплодный (*Populus trichocarpa*), расширило знания о ГБ-содержащих генах растений и позволило классифицировать ГД-содержащие белки по нескольким семействам в соответствии со сходством последовательности гомеодомена, размером белка, расположении ГД в ТФ и по ряду других характеристик [19, 20].

Гомеодомен-содержащие белки можно классифицировать на следующие семейства: PHD, PLINC, WOX, суперкласс TALE (включающий семейства BELL и KNOX), DDT, NDX, LD, PINTOX, SAWADEE и HD-Zip (состоящий из четырех подклассов, I, II, III и IV, соответственно) [4]. Анализ последовательностей ГД организмов различных царств показал, что некоторые ГД-содержащие ТФ растений в большей степени функционально связаны с ТФ, содержащими ГД у животных и грибов, чем с другими ГД растений, принадлежащих к разным семействам. Таким образом, гомеодомены, судя по всему, эволюционно разошлись до момента разделения растений, животных и грибов [19]. Подробная информация о семействах гомеодомен-содержащих белках и их функциях представлена в таблице 1.

1.5 Изучение гомеобокс-содержащих генов растений

Определение возможных функций гомеобокс-содержащих генов на основе результатов анализа дифференциальной экспрессии уже проводи-

Таблица 1 – Классификация и основные функции гомеобокс-содержащих генов модельных организмов [12]

Класс	**Функция**
HD-ZIP I	Реакция на абиотический стресс
HD-ZIP II	Рост и пролиферация клеток
HD-ZIP III	Развитие органов и сосудов
HD-ZIP IV	Дифференциация эпидермальных клеток
KNOX	Инициирование развития и поддержание верхушечной меристемы
BEL	Развитие семяпочки
PLINC	Регуляция цветкового развития
WOX	Определение пути клеточного развития
PHD	Неизвестная функция
DDT	Поддержание вегетативного состояния растения
LD	Регулирование времени цветения
PINTOX	Реакция на некротрофные патогены
SAWADEE	Замалчивание транскрипции путем метилирования ДНК
NDX	Развитие клубеньков

лось для некоторых растений, например для бобовых культур, риса и тканей корня моркови [11, 21, 22]. Также довольно распространены исследования дифференциальной экспрессии данных генов при воздействии на организм растения различных стрессов, например такое исследование проводилось для вида *Brassica rapa* и также бобовых культур [11, 23]. Подобное исследование может представлять интерес для дальнейшего исследования вида *Pinus sylvestris*.

Множество исследований гомеобокс-содержащих генов семейства сосновых посвящено отдельным семействам этих генов. Большее внимание уделено изучению эволюции, экспрессии и функциям генов класса WOX [7–9]. Данные гены являются очень важными в ходе эмбрионального развития организма. У покрытосеменных транскрипционные факторы, принадлежащие семейству WOX, учавствуют регуляции меристемы и в контроле дифференцировки развивающихся эмбрионов [9].

Также имеются исследования других классов, например семейства генов HD-ZIP III, важного для развития древесины [24]. Изучались также и

гены семейства KNOX ели обыкновенной (*Picea abies*) с целью определения их функций в тканях и органах данного вида. Исследователи предположили, что ген *HBK2* учавствует в процессе соматического эмбриогенеза [25].

Однако ранее не проводилось исследований дифференциальной экспрессии гомеобокс-содержащих генов вида *Pinus sylvestris* и выявления их функций в развитии хвои, флоэмы, почки, эмбриона и мегагаметафита (женского гаметофита). Стоит отметить, что эмбриональное развитие *P. sylvestris* уже изучалось, работа была посвящена изучению всех дифференциально экспрессировавшихся транскриптов тканей эмбриона и мегагаметофита, в то время как анализ гомеобокс-содержащих транскриптов не проводился [1].

1.6 Биоинформатический анализ

1.6.1 Анализ данных РНК секвенирования древесных видов растений лесной зоны

Расшифровка последовательности нуклеотидов РНК или РНК секвенирование (РНК-Seq) методами секвенирования следующего поколения (NGS) позволила увеличить число исследований, посвященных экспрессии генов деревьев. Биоинформатический анализ данных РНК-Seq хвойных представляет собой особую сложность по ряду причин, например, из-за огромной длины генома (геномы хвойных находятся в диапазоне от 6 500 до 37 000 миллионов пар нуклеотидных оснований) и небольшого количества хорошо аннотированных геномов [3, 26]. Перечисленные причины означают, что чаще всего для изучения профиля экспрессии генов биоинформатическими методами приходится собирать транскриптом *de novo*. В данном исследовании, посвященном анализу дифференциальной экспрессии генов *P. sylvestris*, необходим собранный транскриптом сосны обыкновенной *de novo*, так как отсутствует референсный геном данного организма. Сборка транскриптома *de novo* означает восстановление последовательностей транскриптов без референсного генома, т.е. без такого генома, который был секвенирован ранее и являлся бы репрезентативны примером набора

генов данного вида.

В настоящее время проводится всё больше исследований, посвящённых количественной оценке дифференциальной экспрессии матричной РНК различных органов или тканей деревьев на основе данных РНК секвенирования. Более того, изучение дифференциальной экспрессии применяется при изучении реакций организма в различных условиях среды. Наиболее изученные виды среди группы голосемянных относятся к семейству сосновых, а в группе покрытосеменных к семейству ивовых, розоцветных, бобовых и буковых (рисунок 4). РНК секвенирование является хорошим способом улучшения результатов сборки и аннотирования геномов [27, 28].

1.6.2 Данные РНК секвенирования

Качество взятых из открытого доступа данных РНК-секвенирования необходимо проверить и отобрать только те, которые соответствуют установленным критериям. При подготовке библиотек РНК используются специальные методы для очистки данных от загрязнения рибосомальной РНК (рРНК), однако в конечных данных могут встречаться рРНК и поэтому необходима дополнительная фильтрация. Рибосомальная РНК не представляет интереса для изучения, хотя она и составляет большую часть РНК в клетке. Исследователя интересует матричная РНК, кодирующая часть которой определяет аминокислотные последовательности белков.

1.6.3 Сборка транскриптома *de novo*

Транскриптом — это множество транскриптов, а транскрипты в свою очередь представляют собой какой-либо ген исследуемого организма, в данном случае сосны обыкновенной. Программа Trinity объединяет три независимых программных модуля: Inchworm, Chrysalis и Butterfly, последовательно применяемых для обработки больших объемов последовательностей RNA-seq. Trinity разделяет данные последовательности на множество отдельных графов де Брейна, каждый из которых представляет собой множество вариантов транскрипта для данного гена или локуса, затем програм-

ма обрабатывает каждый граф независимо, с целью извлечения изоформ сплайсинга и выделения отдельных транскриптов, полученных из паралогичных генов.

- Inchworm собирает данные RNA-seq в уникальные последовательности транскриптов;
- Chrysalis кластерезует последовательности, полученные на предыдущем шаге и строит полные графы де Брейна для каждого кластера;
- Butterfly выделяет полноразмерные транскрипты для альтернативно сплайсированных изоформ и также те транскрипты, которые соответствуют паралогичным генам. Иными словами, модуль выводит все возможные последовательности транскриптов [29].

1.6.4 Дифференциальная экспрессия генов

Экспрессия генов — процесс, с помощью которого информация, содержащаяся в гене, используется в синтезе функционального продукта гена — белка. Ген называется дифференциально экспрессированным, если наблюдаемая разница в изменении количества прочтений (РНК) или уровней экспрессии между двумя экспериментальными условиями является статистически значимой [30].

Для количественной оценки экспрессии генов данные РНК секвенирования (прочтения) выравниваются на эталонный геном, если таковой имеется, однако чаще всего в случае отсутствия референсного генома выравнивание осуществляется на транскриптом, собранный *de novo*.

Существует множество программ, для осуществления данной задачи, например программа RSEM. В данной программе оценивается количество выровненных прочтений для каждого транскрипта. Результатом работы на будут являться файлы, уникальные для различных образцов. Каждый такой файл содержит данные об уровне экспрессии для каждого транскрипта, а также колонки «TPM» (Transcripts per million) и «FPKM» (Fragments per kilobase of cDNA per million fragments mapped), содержащие нормированные уровни экспрессии для каждого транскрипта.

Число выровнявшихся прочтений используется для оценки относительного уровня экспрессии генов, а затем с помощью различных методов нормализации проверяется значимость различий между группами образцов [3]. Примером таких методов может служить (TC) нормализация, где количество ридов, выровнявшихся на транскрипты, делится на общее количеству ридов. Другой метод, используемый в данной работе, TMM (The trimmed mean of M values) нормализация [3]. Суть данного метода заключается в использовании взвешенного усеченного среднего значения логарифмических отношений количества выровнявшихся ридов двух образцов. В данном методе используется предположение, что большая часть генов не экспрессируется дифференциально [31].

В данной работе для определения роли гомеобокс-содержащих генов в различных тканях и на различных этапах развития сосны обыкновенной проанализированы дифференциально экспрессирующиеся гомеобокс-содержащие гены для эмбриона, женского гаметофита, хвои, почки и флоэмы, соответственно.

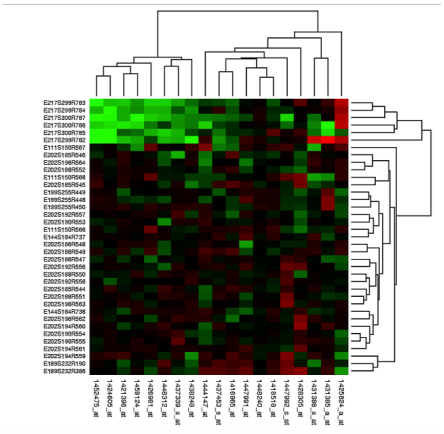

Рисунок 3 – Кластерная тепловая карта [32]

Для наглядной визуализации результатов строится тепловая карта. В данной работе используется одна из её разновидностей — кластерная тепловая карта. Данный объект является графическим изображением, представляющим собой иерархическую структуру кластеров строк и столбцов матрицы данных. Кластерная тепловая карта состоит из прямоугольного мозаичного изображения, причем каждый элемент данного мозаичного изображения окрашен по определённой цветовой шкале с целью представления значения соответствующего элемента матрицы данных. Строки (столбцы) упорядочены таким образом, что похожие строки (столбцы) располагаются рядом друг с другом. На вертикальных и горизонтальных полях мозаичного изображения расположены иерархические кластерные деревья [32]. Пример подоб-

ной тепловой карты можно увидеть на Рисунке 3.

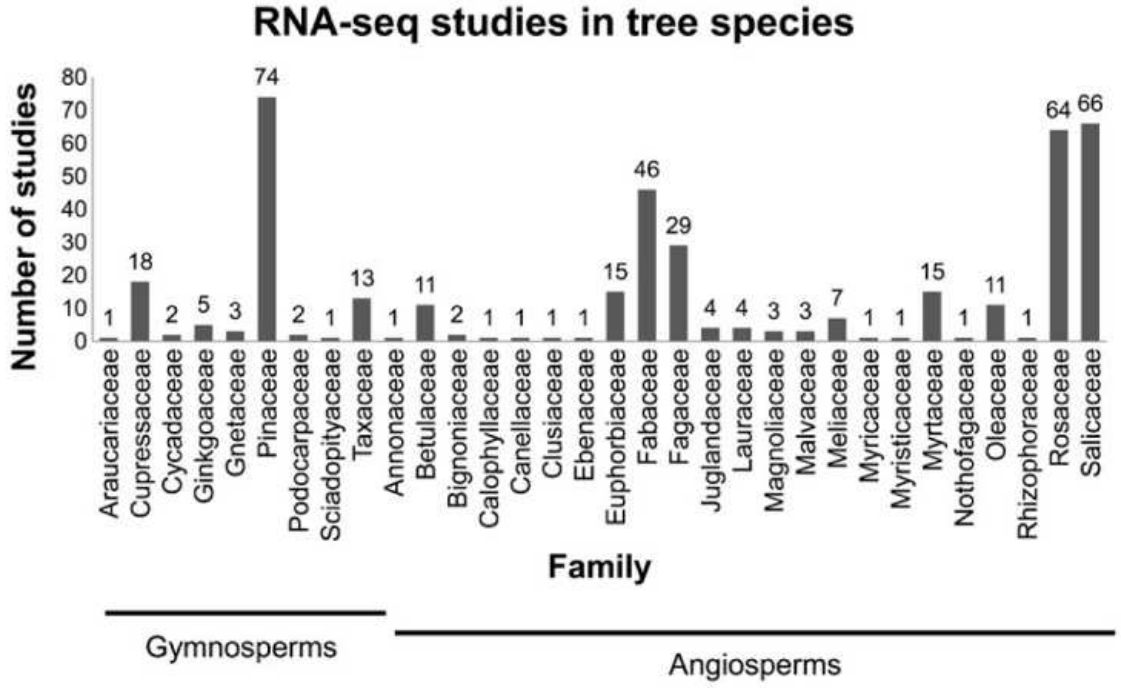

Рисунок 4 – Исследования различных деревьев на основе РНК секвенирования (классификация видов по семействам), где gymnosperms — голосеменные, angiosperms — покрытосемянные [3]

2 Материалы и методы

Материалом для данной работы послужили открытые данные РНК-секвенирования сосны обыкновенной (*Pinus sylvestris*), размещенные в базе данных NCBI BioProject под номером доступа PRJNA531617 [33]. Данные предоставлены Норвежским институтом биоэкономических исследований (Norwegian Institute of Bioeconomy Research). Сами последовательности или риды, которые являются результатом РНК-секвенирования представляют собой нуклеотидные последовательности длиной около 150 н. о. Материал содержит прочтения пяти тканей сосны обыкновенной (хвоя, флоэма, вегетативная почка, эмбрион и женский гаметофит), выделенный из шести деревьев сосны обыкновенной, выросших в лесу рядом с муниципалитетом Пункахарью, расположенным в южной Финляндии (ныне являющийся частью муниципалитета Савонлинна). Сбор производился 26-27 мая 2016 года.

Для обработки данных, сборки транскриптома и анализа транскриптов был использован специализированный комплекс IBM для высокопроизводительных вычислений с 96 ядерным SMP сервером IBM x3950 X6 с объемом ОЗУ 3 ТБ. Также в комплекс входит гибридный счетный сервер IBM dx360 M4 с двумя GPU NVIDIA Tesla K20, обеспечивающий суммарную пиковую производительность для вычислений одинарной точности с плавающей точкой до 7 Tflops, а также подсистема хранения данных IBM Storwize V3700 объемом 48 Тб. Комплекс работает под управлением ОС SuSe Linux Enterprise Server 11, установлена параллельная файловая система IBM GPFS, система мониторинга Ganglia и система пакетной обработки Torque.

2.1 Предобработка данных и *de novo* сборка транскриптома *P. sylvestris*

Для очищения ридов от фракции рибосомальной РНК использовалась программа SortMeRNA версии 2.1 и библиотека Silva, содержащая

информацию о нуклеотидных последовательностях малой и большой субъединиц рРНК для архей, бактерий и эукариот.

Программа FastQC использовалась для первичной оценки качества данных РНК секвенирования. По результатам анализа данной программы подбирались дальнейшие параметры для фильтрации прочтений [34]. Устранение некоторых последовательностей и обрезка участков ридов с низким качеством производилась с помощью программы Trimmomatic (версия 0.33) [35]. Для этого использовались следующие параметры программы: «ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:1:true TRAILING:3 SLIDING WINDOW:4:25 AVGUAL:30», где

- ILLUMINACLIP — шаг, использующийся для поиска и удаления адаптеров Illumina;
- TruSeq3-PE-2.fa — библиотека адаптеров. На начальном этапе, Trimmomatic ищет начальные совпадения (16 пар нуклеотидов), позволяющие максимум **2** несоответствия. Если прочтения являются парноконцевыми, то при достижении оценки качества **30**, или в случае одноконцевых прочтений оценки качества **10**, участки последовательностей, удовлетворяющие данным значениям качества, будут расширены и обрезаны. Параметр **true** позволяет сохранить обратное чтение;
- TRAILING — команда, обрезающая нуклеотиды с конца прочтения, в случае если качество ниже порогового, то есть меньше **3**;
- SLIDINGWINDOW — данная команда единожды обрезает нуклеотиды для каждого прочтения, если среднее качество в пределах перемещающегося окна ниже порогового значения, **4** — размер окна, **25** — качество;
- AVGUAL — шаг, удаляющий прочтение из множества, если среднее значение качества ниже **30**.

Для сборки транскриптома de novo использовалась программа Trinity (версия 2.8.4) [29].

2.2 Оценка полноты сборки транскриптома

Оценка полноты сборки транскриптома осуществлялась с помощью программы BUSCO (Benchmarking Universal Single-Copy Orthologs) (версия 3) и базы данных однокопийных ортологов embryophyta odb9 [36]. Ортологи являются гомологичными генами, обнаруженными у разных видов или внутри одного организма. Результат работы программы представляет собой шесть параметров, являющиеся метрикой законченности и полноты транскриптома, C:complete [S:Single, D:duplicated], F:fragmented, M:missed, n:number of genes. Первые пять параметров отражают процент обнаруженных генов из базы данных однокопийных ортологов BUSCO. Идентифицированные гены ортологов могут быть полными и представленными единичной копией (Completed Single), полными и дублированными (Completed Duplicated), фрагментарными (Fragmented) и пропущенными (Missed). Последний параметр n характеризует количество используемых для оценки генов [36]. Хорошее качество сборки транскриптома характеризуется относительно высокими значениями параметров Completed, однако нет единого соглашения, начиная с какой величины следует считать значения параметра высокими. Чем ближе C к 100 %, тем более полным является транскриптом.

2.3 Отбор гомеобокс-содержащих транскриптов

Сосна обыкновенная является немодельным организмом, однако так как гомеодомен является весьма консервативным, его можно обнаружить и у данного вида. Программа HMMER (версия 3.2.1) использовалась для поиска и идентификации гомеодоменов в транскриптоме *P. sylvestris* [37, 38]. Программа использует скрытую марковскую модель (CMM) гомеодомена, загруженную из базы данных семейств белковых доменов PFAM (номер модели PF00046) [39].

2.4 Анализ дифференциальной экспрессии генов и построение тепловой карты

Анализ дифференциальной экспрессии является многоступенчатым процессом; для выявления профилей дифференциальной экспрессии гомеобокс-содержащих транскриптов использовались такие программы как Trinity, RSEM (версия 1.3.2), EdgeR (версия R 3.5.0, версия Bioconductor 3.8) [40]. Для оценки уровня экспрессии транскриптов использовался скрипт align_and_estimate_abundance.pl, входящий в пакет Trinity, а также программы RSEM и Bowtie (версия 1.2.3). Bowtie выравнивает отфильтрованные по качеству прочтения, полученные в результате РНК-секвенирования на собранные транскрипты. Позже, с помощью программы RSEM оценивается количество выровненных прочтений для каждого транскрипта [41]. Поскольку количество транскриптов может существенно различаться между различными образцами, прочтения, полученные из каждого образца, необходимо выравнивать на транскрипты независимо, в итоге получая специфичные для каждого образца значения количества выровнявшихся ридов.

Следующий шаг представляет собой объединение всех данных об уровне экспрессии каждого транскрипта для различных образцов. Эта процедура осуществляется с помощью встроенного в Trinity скрипта abundance_estimates_to_matrix.pl и также программы RSEM. В результате получаются две матрицы, одна из которых содержит информацию о количестве прочтений, полученных для транскрипта для каждого отдельного образца, а другая — значения уровней экспрессии, дополнительно нормализованные между образцами с использованием метода нормировки TMM для корректировки любых различий в составе образца [31]. Данная нормализованная матрица используется в дальнейшем для анализа дифференциальной экспрессии.

В настоящее время доступно множество инструментов для идентификации дифференциально экспрессируемых транскриптов, и программа edgeR является одной из наиболее популярных и точных [40]. Программ-

ное обеспечение edgeR является частью пакета R Bioconductor, который в свою очередь используется во встроенном скрипте run_DE_analysis.pl программы Trinity для анализа дифференциальной экспрессии. Этот шаг EdgeR генерирует матрицы, содержащие результаты всех парных сравнений между пятью образцами (хвоя и почка, хвоя и мегагаметофит и т. д.). Полученные матрицы включают в себя значения logFC, logCPM (log counts per million), p-value для каждого теста и FDR (false discovery rate). logFC — двоичный логарифм отношения уровней экспрессии для образцов разных тканей. Положительные значения logFC указывают на высокую экспрессию транскрипта, а отрицательные на пониженную экспрессию.

Благодаря скрипту Trinity analyze_diff_expr.pl осуществляется финальный шаг в определении дифференциально экспрессирующихся транскриптов и построении тепловой карты. Для работы данного скрипта необходимо указать два параметра, p-значение и logFC. В настоящем исследовании p-значение равно 0.001, параметр C — 3. Параметр C характеризует изменение в экспрессии транскриптов, в данном случае, например, будет 2^3 или 8-кратное изменение.

С помощью иерархической кластеризации генерируется тепловая карта, отображающая кластеризацию одинаково экспрессируемых транскриптов (вертикальная ось), отнесенные к типу образца (горизонтальная ось). Порядок столбцов соответствует чередованию образцов [42]. Для построения тепловой карты отбираются только те транскрипты, которые хотя бы в одном попарном сравнении между образцами отличаются в уровне экспрессии в 8 раз. Численные значения экспрессии визуализированы в пространстве log2, также среднее значение экспрессии (значения TMM-нормализованной матрицы экспрессии) для каждого транскрипта вычитается из значений экспрессии для каждого значения в строке. Желтым цветом выделены образцы с повышенной экспрессией, фиолетовым — с пониженной.

2.5 Аннотация дифференциально экспрессирующихся транскриптов

Программа OmicsBox (версия 1.2.4) использовалась для аннотации дифференциально экспрессирующихся гомеобокс-содержащих транскриптов. Для каждого транскрипта с помощью встроенного алгоритма действий (Gene Onthology Annotation workflow) были определены гены и белки, кодируемые данными генами. Для этой задачи использовался алгоритм BLAST и открытая база данных последовательностей белков SwissProt/UniProt [43, 44]. Для каждого гена был определены GO термы (terms). «Генная онтология» (Gene Ontology (GO)) представляет собой унифицированную терминологию для аннотации генов и генных продуктов, база данных генной онтологии включает в себя три независимых словаря — это молекулярные, клеточные и биологические функции. Генная онтология описывает комплексные биологические феномены, а не конкретные биологические объекты [45].

3 Результаты и обсуждение

3.1 Сборка транскриптома и оценка качества

Собранный траскриптом *P. sylvestris* содержит 775 502 транскрипта со средним содержанием GC-контента 40.19%. Значение N50 равно 1 273 н. о., а медианное значение длины транскрипта 360 н. о. (таблица 2). Полнота сборки оценена с помощью 6 параметров BUSCO, в результате были получены следующие параметры C:85.9% [S:27.3%, D:58.6%] F:2.2%, M:11.9%, n:1440 (C:complete [S:Single, D:duplicated], F:fragmented, M:missed, n:number of genes) (Рисунок 5). Данные результаты свидетельствуют о более полной и законченной сборки транскриптома сосны обыкновенной, по сравнению с предыдущей сборкой [33]., выполненной группой исследователей из Финляндии (Biocenter Oulu, University of Oulu , Norwegian Institute of Bioeconomy Research, Flanders Research Institute for Agriculture, Natural Resources Institute Finland, University of Helsinki) в 2019 году. В данной работе получены следующие значения параметров BUSCO: C:71.9% [S:26.3%, D:45.6%] F:8.7%, M:19.4%, n:1440.

3.2 Классификация гомеобокс-содержащих транскриптов

С помощью программы HMMER было отобрано 243 уникальных гомеобокс-содержащих транскрипта. Используя программу OmicsBox, данные транскрипты были проаннотированы (E-value $< 10^{-3}$). Ожидаемое значение Expected value (E-value) — это параметр, который описывает количество совпадений, которые можно обнаружить «случайно» при поиске в базе данных определенного размера. В сущности, E-value описывает случайный фоновый шум. Чем ниже E-value или чем оно ближе к нулю, тем более значимым является совпадение.

Транскрипты классифицированы по 14 классам генов, содержащих гомеобокс (Таблица 3). Наиболее представленными классами генов, содержащих гомеобокс, оказались HD-ZIP I, HD-ZIP II, WOX, и HD-ZIP IV. Гомеодомены класса HD-ZIP I участвуют в реакциях, связанных с реакцией

Таблица 2 – Статистика сборки транскриптома *P. sylvestris*, где в колонке $Trinity_{\text{СФУ}}$ указаны результаты сборки, полученной в данной работе, $Trinity_{\text{Финляндия}}$ — результаты сборки, полученной группой финских исследователей.

Группа сборки	$Trinity_{\text{СФУ}}$	$Trinity_{\text{Финляндия}}$
Общее количество «генов» Trinity	488 116	—
Общее количество транскриптов	775 502	1 288 196
GC%	40.19	40.39
Статистика по всем транскриптам		
Значение N50 (н. о.)	1 273	658
Медианное значение длины транскрипта (н. о.)	360	320
Средняя длина транскрипта (н. о.)	713.6	529.4
Общее количество оснований (н. о.)	553 398 248	681 954 381
Статистика по самым длинным изоформам «генов» транскриптов		
Значение N50 (н. о.)	596	—
Медианное значение длины транскрипта (н. о.)	313	—
Средняя длина транскрипта (н. о.)	513.56	—
Общее количество оснований (н. о.)	250 675 880	—

организма на абиотический стресс, абсцизовую кислоту, синий свет, а также играют роль процессах деэтиоляции и эмбриогенеза [46]. Гомеодомены класса HD-ZIP II известны своей ролью в реакциях избегания растением тени, регуляции развития верхушечной (апикальной) части эмбриона и функций меристемы [47]. Известно, что гомеодомены класса WOX являются неотъемлемой частью многочисленных процессов развития, а именно, формирования паттернов эмбриона, органов, сохранения стволовых клеток. Белки, кодируемые генами класса HD-ZIP IV участвуют в процессах репродуктивного развития, развития проростка, сохранения эпидермального клеточного слоя [48, 49].

Следует отметить, что некоторые из транскриптов были неклассифицированны, что в свою очередь может представлять интерес для дальнейшего исследования функций данных транскриптов.

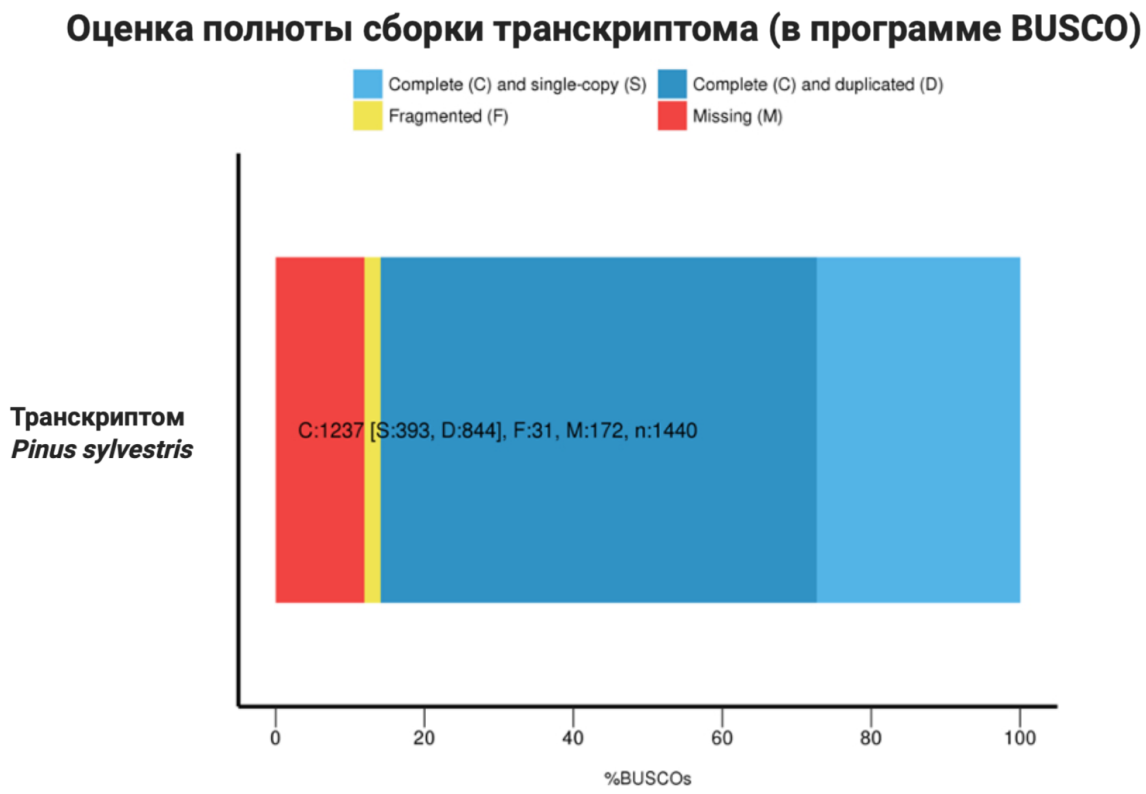

Рисунок 5 – Характеристика полноты сборки транскриптома *P. sylvestris* по результатам работы программы BUSCO

Таблица 3 – Распределение гомеобокс-содержащих транскриптов по классам генов, содержащих гомеобокс, где **N** — количество транскриптов, **U** — неклассифицированные, **T** — общее количество.

Класс	**N**	Класс	**N**	Класс	**N**	Класс	**N**
HD-ZIP I	67	PLINC	1	DDT	11	NDX	0
HD-ZIP II	29	WOX	29	PHD	4	SAWADEE	11
HD-ZIP III	15	BEL	21	PINTOX	0	**U**	14
HD-ZIP IV	26	KNOX	11	LD	4	**T**	243

3.3 Дифференциальная экспрессия гомеобокс-содержащих генов в различных тканях *Pinus sylvestris*

Дифференциально экспрессирующиеся гомеобокс-содержащие гены визуализированы в виде тепловой карты (Рисунок 7) и также представлены в виде графической иллюстрации (Рисунок 6). Идентифицировано 46 статистически значимо дифференциально экспрессировавшихся «генов» Trinity.

Рассмотрим общий профиль экспрессии генов эмбриона и мегагаметофита. На ранних стадиях развития в эмбрионе и мегагаметофите повышена экспрессия генов *WOX11*, *WOX11*, *ROC1*, *ROC2*, *ROC8*, *PDF2*, *HDG2*, *HAT14* и гена, кодирующего белок позднего эмбриогенеза D-34 (далее в тексте просто *X*). Понижена экспрессия генов *HOX5*, *HAT7*, *HOX21*, *ATHB-23*, *WUS*. Можно увидеть, что паттерны экспрессии генов весьма схожи для данных тканей. Рассмотрев совместную повышенную и пони-

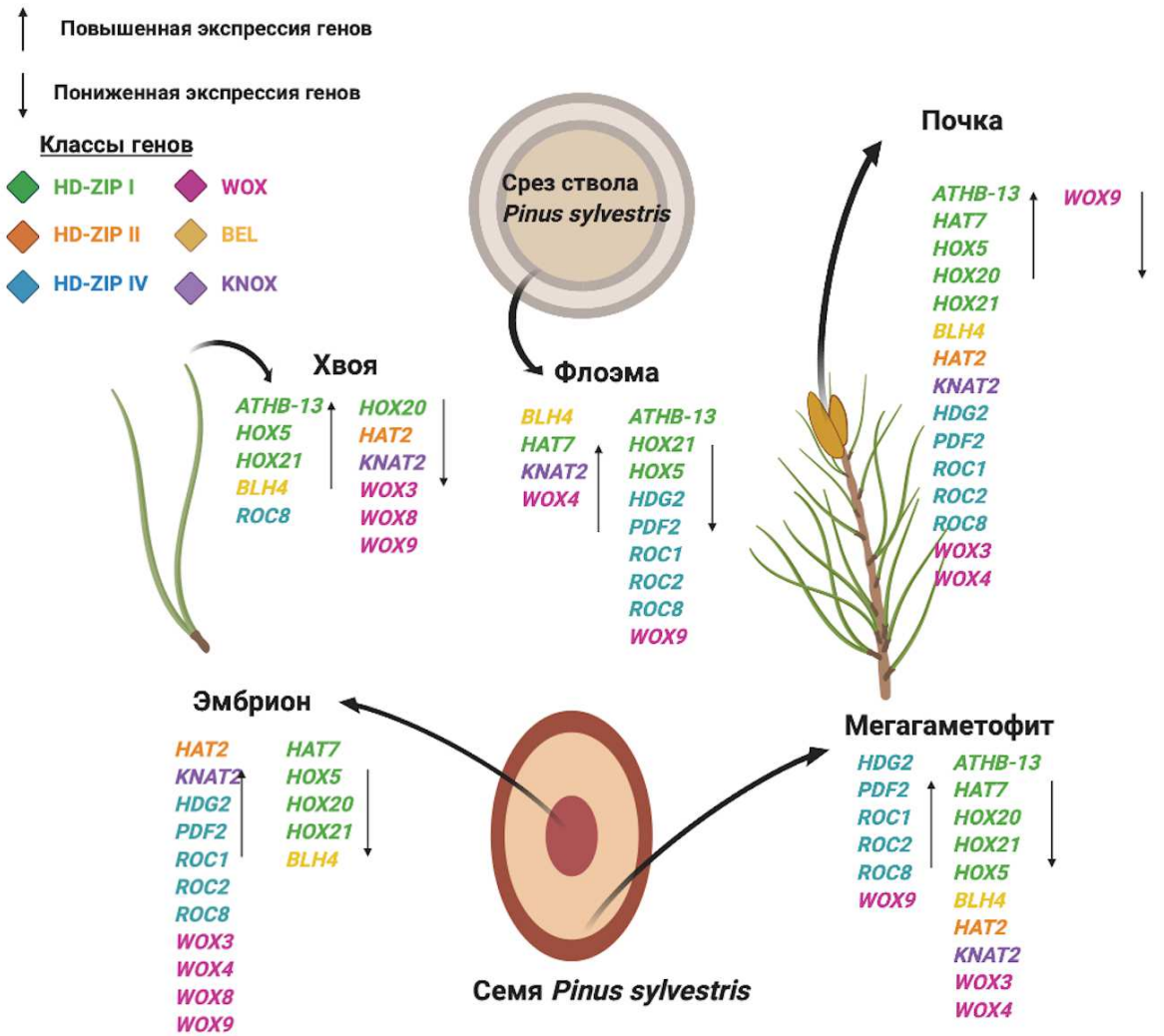

Рисунок 6 – Графическая иллюстрация дифференциально-экспрессировавшихся генов *P. sylvestris*.

женную экспрессию генов в тканях эмбриона и мегагаметофита, перейдем к рассмотрению различающихся профилей экспрессии для двух данных тканей. В эмбриональной ткани также повышена экспрессия генов *HAT2*, *HOX20*, *KN-1*, *KNAT2*, *WOX3*, *WOX4*, *WOX8* и понижена экспрессия генов

ATHB-5, *BLH1*, *HAT14*.

В ткани мегагаметофита понижена экспрессия генов *ATHB-13*, *ATHB-23*, *HAT2*, *HOX20*, *KN-1*, *KNAT1*, *KNAT2*, *KNAT3*, *ZHD1*, *WOX3* и *WOX4*. Ген *HDG2* с повышенной экспрессией обнаружен только в тканях женского гаметофита, эмбриона и почки, что может говорить о вероятной специальной функции белка, участвующего в развитии данных молодых тканей и органов. Анализируя вышеописанные паттерны экспрессии генов, можно предположить, что гены *HAT2*, *HOX20*, *KN-1*, *KNAT2*, *WOX3*, *WOX4*, *WOX8* учувствуют в развитии эмбриона. Ранее было показано, что гены *WOX2*, *WOX8* и *WOX9* обладают повышенной экспрессией в эмбрионе *Picea abies* и *Arabidopsis thaliana* [8, 9, 50], также обнаружено, что экспрессия гена *WOX3* также повышена в эмбриональной ткани *Pinus pinaster* [51]. Имеются данные, что на эмбриональной стадии развития сосны обыкновенной активен ген *HAT5* [1].

Рассмотрим дифференциально экспрессирующиеся гомеобокс-содержащие гены хвои. В этой ткани были обнаружены гены с повышенной экспрессией, такие как *ATHB-13*, *BLH1*, *BLH4*, *HAT14*, *HOX5*, *HOX21*, *KNAT3*, *ROC8*, *ZHD1* и также гены с пониженной экспрессией, а именно *X*, *HAT14*, *HAT2*, *HOX20*, *KN-1*, *KNAT1*, *KNAT2*, *WOX3*, *WOX8*, *WOX9*, *WOX11*. Ранее было показано, что экспрессия генов WOX9 и WOX11 понижена для листьев риса [6].

В ткани почки наблюдалась повышенная экспрессия большинства гомеобокс-содержащих генов (см. Таблица 4) и пониженная экспрессия генов *X*, *WOX9*, *WOX11*. В почке активно идут процессы морфогенеза, о чем свидетельствует повышенная экспрессия отвечающих за развитие и рост генов. Показано, что ген *WOX3* играет важную роль в развитии боковых органов (листья и подобные им органы) [51, 52]. Ген *WUS* экспрессируется в верхушечных меристемах, однако о его роли в развитии почки *Pinus sylvestris* пока ничего не известно [51, 53]. Гены *HOX5*, *HOX20*, *HOX21*, *HAT7* принадлежат к классу генов HD-ZIP I и их функции внутри группы голосемянных растений еще очень мало изучены. Гены данного класса чаще всего учувствуют в реакциях растения на стресс и регуляции

Таблица 4 – Дифференциально экспрессирующиеся гомеобокс-содержащие гены пяти тканей *Pinus sylvestris*, где ЭМ — эмбрион, МГ — мегагаметофит, ХВ — хвоя, ПЧ — почка, ФЛ — флоэма.

Образец	ЭМ	МГ	ХВ	ПЧ	ФЛ
Ген/Класс	HD-ZIP I				
ATHB-23	↓	↓	—	—	—
HAT7	↓	↓	—	↑	↑
HOX5	↓	↓	↑	↑	↓
HOX20	—	↓	↓	↑	—
HOX21	↓	↓	↑	↑	↓
	HD-ZIP II				
HAT2	↑	↓	↓	↑	—
HAT14	—	↑	—	↑	↑
	HD-ZIP IV				
HDG2	↑	↑	—	↑	↓
PDF2	↑	↑	—	↑	↓
ROC1	↑	↑	—	↑	↓
ROC2	↑	↑	—	↑	↓
ROC8	↑	↑	↑	↑	↓
	PLINC				
ZHD1	↑	↓	↑	↑	—
	WOX				
WOX3	↑	↓	↓	↑	—
WOX4	↑	↓	—	↑	↑
WOX8	↑	—	↓	—	↑
WOX9	↑	↑	↓	↓	↓
WOX11	↑	↑	↓	↓	↓
WUS	↓	↓	—	↑	—
	BEL				
BLH1	↓	—	↑	—	↑
BLH4	↓	↓	↓	↑	↑
	KNOX				
KN-1	↑	↓	↓	↑	↑
KNAT1	—	↓	↓	↑	↑
KNAT2	↑	↓	↓	↑	↑
KNAT3	—	↓	↑	↑	↑

развития органов растений [54]. Судя по всему, данные гены также принимают участие в подобных реакциях в почке и хвое сосны обыкновенной. Также повышенной экспрессией обладают гены *HAT14*, *HAT2*, принадлежащие классу HD-ZIP II.

В предыдущих исследованиях установлено, что гены данного класса учувствуют в процессах, связанных с реакцией на свет и сигнализацией ауксина [54–58], можно предположить, что ген *HAT2* с повышенной регуляцией характерен только для почки и эмбриона и выполняет какую-то специфичную функцию в их развитии, в то время как ген *HAT14* экспрессируется во всех пяти тканях сосны обыкновенной. В ткани почки выявлен класс генов HD-ZIP IV, а именно гены *PDF2*, *ROC1*, *ROC2*, *ROC8*. Гены данного класса участвуют в развитии кутикулы растений, аккумуляции антоцианов и дифференцировки эпидермальных клеток [19, 54, 59]. Возможно, данные гены выполняют схожие функции в ткани почки и эмбриона. Недавние исследования показывают, что ген *BLH4* участвует в реакции диметилэстерификации (этерификация— реакция образования сложных эфиров при взаимодействии кислот и спиртов) гомогалактуронана, что определяет структуру и функции полисахарида пектина, важнейшего компонента клеточной стенки [60].

Таблица 5 – Распределение дифференциально экспрессирующихся генов по классам, также сюда входит один неклассифицированный ген *X* (gene encoding LEA).

Класс	HD-ZIP I	HD-ZIP II	HD-ZIP IV	PLINC	WOX	BEL	KNOX
Гены	*HAT7*	*HAT14*	*HDG2*	*ZHD1*	*WOX3*	*BLH1*	*KN-1*
	ATHB-13	*HAT2*	*PDF2*		*WOX4*	*BLH4*	*KNAT1*
	ATHB-23		*ROC1*		*WOX8*		*KNAT2*
	HOX5		*ROC2*		*WOX9*		*KNAT3*
	HOX20		*ROC8*		*WOX11*		
	HOX21				*WUS*		

В последней из рассматриваемых тканей, а именно во флоэме, наблюдается повышенная экспрессия генов *BLH1*, *BLH4*, *HAT14*, *HAT7*, *KN-1*, *KNAT1*, *KNAT2*, *KNAT3*, *WOX4*, *WOX8*. Пониженной экспрессией об-

ладают гены *X*, *ATHB-13*, *HDG2*, *HOX5*, *HOX21*, *PDF2*, *ROC1*, *ROC2*, *ROC8*, *WOX9* и *WOX11*. Ген *WOX4* участвует в развитии сосудистых тканей, в дифференцировки камбия и развитии прокамбия, который является первичной сосудистой меристемой, которая затем дифференцируется в первичную ксилему и флоэму, эти результаты были получены для деревьев рода *Populus* и для видов *Pinus pinaster*, *Solanum lycopersicum*, *Arabidopsis* [51, 61, 62]. Для группы голосемянных растений ранее не описывалась повышенная экспрессия и функции гена *WOX4* в ткани флоэмы, являющейся сосудистой тканью. По полученным результатам можно предположить, что данный ген также характерен для развития флоэмы в зрелых тканях *Pinus sylvestris*.

Следует отметить, что наибольшее число генов класса *KNOX*, а именно *KN-1*, *KNAT1*, *KNAT2*, *KNAT3*, с повышенной экспрессией обнаружены в тканях флоэмы и почки. Недавние исследования показывают, что ген *KNAT2/6b*, ортолог генов *KNAT2* и *KNAT6* у *Arabidopsis*, высоко экспрессирован в тканях флоэмы и ксилемы тополя [63]. Ген *KNAT3* принадлежит к классу II KNOX; гены этого класса известны своей ролью в регуляции развития вторичной клеточной стенки [64]. Во флоэме и почке также повышена экспрессия генов класса BEL — это *BLH1*, *BLH4* соответственно. Гены класса KNOX и BEL взаимодействуют между собой и учувствуют в развитии верхушечной меристемы и развитии органов плодов и соцветий у *Arabidopsis* [64,65]. Таким образом, вероятно во флоэме и почке тоже имеет место взаимодействие классов генов KNOX и BEL и возможно ген *KNAT2* участвует в развитии проводящей ткани. Для уточнения возможных взаимодействий этих генов требуется дальнейший анализ.

Вышеописанные гены с пониженной и повышенной экспрессией в тканях сосны обыкновенной представлены в таблице 4 и таблице 5. Результаты приведенного анализа могут служить отправной точкой для дальнейших исследований функций гомеобокс-содержащих генов в других тканях и органах сосны обыкновенной, с дальнейшей целью улучшения качеств древесины, устойчивости к абиотическим и биотическим стрессам и т. д. В пер-

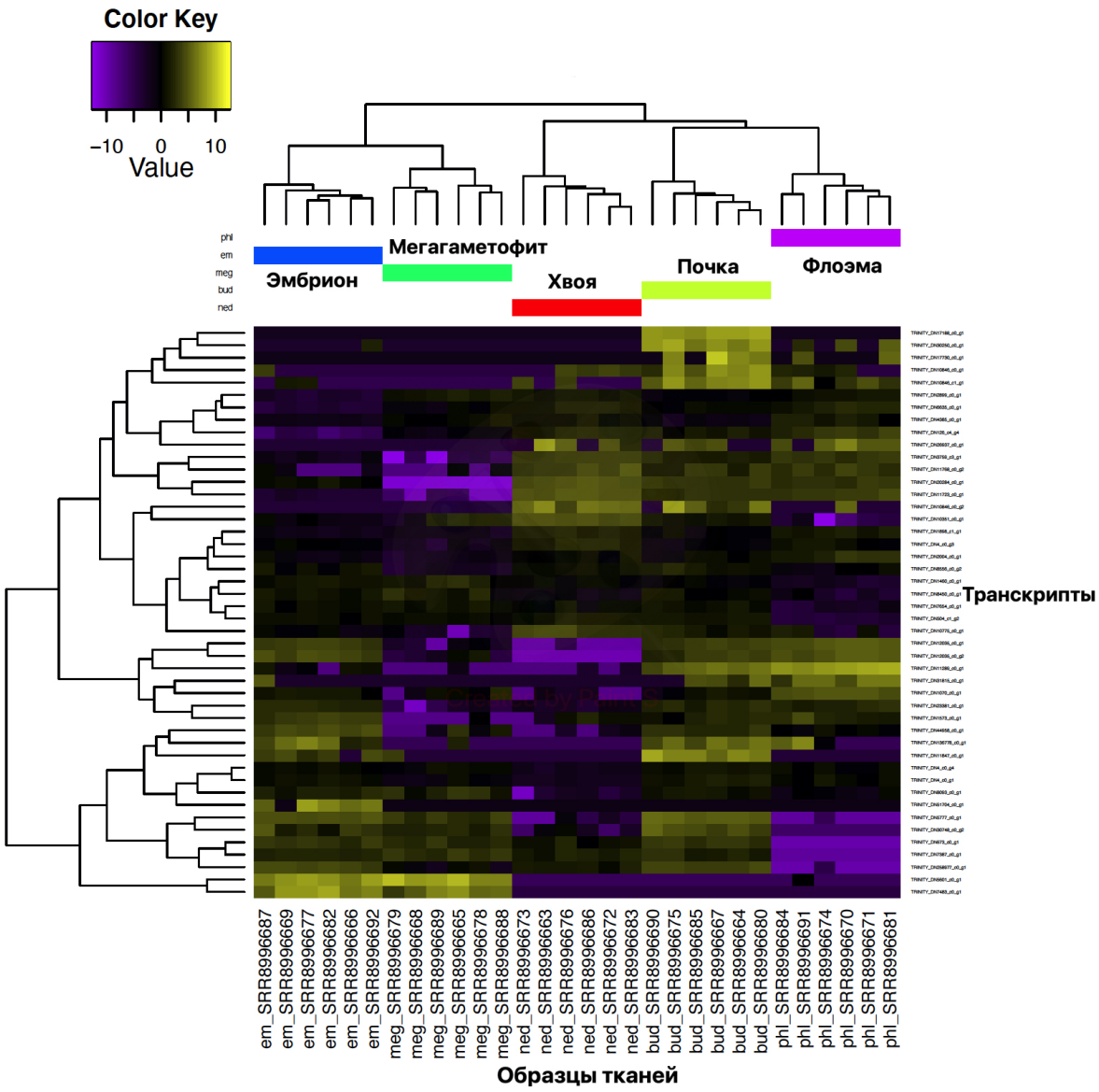

Рисунок 7 – Тепловая карта экспрессии генов пяти тканей *P. sylvestris*.

спективы исследования входит изучение реакции сосны обыкновенной на различные виды стресса, другими словами, исследование изменения экспрессии гомеобокс-содержащих генов в ответ на раздражающий фактор. Подобный анализ мог бы быть полезным лесному хозяйству.

ЗАКЛЮЧЕНИЕ

В ходе данной работы были выполнены все задачи, и поставленная цель была полностью достигнута:

1. Произведен отбор данных РНК-секвенирования различных тканей сосны обыкновенной из базы NCBI;
2. Выполнена предобработка данных и оценка качества ридов;
3. Собран *de novo* транскриптом *Pinus sylvestris*, содержащий 775 502 транскрипта общая длина которого составила 553 398 248 со средним содержанием GC-контента 40.19 %. сборка $Trinity_{\text{СФУ}}$ по результатам анализа является более полной, чем сборка $Trinity_{\text{Финляндия}}$
4. Отобрано и аннотировано 243 гомеобокс-содержащих транскрипта;
5. Определены дифференциально экспрессирующиеся гомеобокс-содержащие гены пяти различных тканей сосны обыкновенной и построена тепловая карта (рисунок 5);
6. Предположительно определены функции некоторых гомеобокс-содержащих генов:
   - Гены *HDG2*, *PDF2*, *ROC1*, *ROC2*, *ROC8* класса HD-ZIP IV участвуют в развитии кутикулы растений, аккумуляции антоцианов и дифференцировки эпидермальных клеток тканей почки и эмбриона; ген *ROC8* также участвует в дифференцировке эпидермальных клеток хвои;
   - Ген *WOX3* семейства WOX экспрессируется в почках и играет роль в развитии боковых органов сосны обыкновенной;
   - Ген *WOX4* характерен для развития сосудистых тканей *Pinus sylvestris*, в том числе и флоэмы;
   - Гены *WOX8* и *WOX9* участвуют в развитии эмбриона *Pinus sylvestris*;
   - Ген *HAT2* экспрессируется в почке и эмбрионе и участвует в процессах, связанных с реакцией на свет и сигнализацией ауксина;
   - Ген *BLH4* участвует в формировании клеточных стенок хвои, почки

и флоэмы;

- Во флоэме и почке имеет место взаимодействие классов генов KNOX и BEL и вероятно ген *KNAT2* участвует в развитии проводящей ткани, гены *KN-1*, *KNAT1*, *KNAT3* имеют аналогичные функции в тканях эмбриона, почки и флоэмы;
- Гены *HOX5*, *HOX20*, *HOX21*, *HAT7* принадлежат к классу генов HD-ZIP I учувствуют в реакциях организма на стресс и регуляции развития почек сосны обыкновенной;
- Гены *ATHB-13*, *PDF2*, *ROC1*, *ROC2*, *ROC8* класса HD-ZIP IV участвуют в развитии кутикулы растений, аккумуляции антоцианов и дифференцировки эпидермальных клеток тканей почки и эмбриона.

Подводя итог, можно сказать, что была получена более качественная сборка транскриптома сосны обыкновенной (775 502 транскрипта, N50 1 273 н.о., BUSCO C:85.9 %) по сравнению со сборкой, полученной финскими коллегами ($Trinity_{\text{Финляндия}}$) и впервые описаны и предположительно определены функции гомеобокс-содержащих генов в пяти тканях *P. sylvestris*.

СПИСОК СОКРАЩЕНИЙ

1. ДНК — дезоксирибонуклеиновая кислота;
2. РНК — рибонуклеиновая кислота;
3. РНК-сек — РНК секвенирование;
4. ГД — гомеодомен;
5. ГБ — гомеобокс;
6. ТФ — транскрипционный фактор;
7. н. о. — нуклеотидные основания;
8. GC состав (гуанин-цитозин состав) — процентный состав суммы всех гуанинов (G) и цитозинов (C) по отношению к длине исследуемого участка нуклеиновых кислот;
9. TC —Total counts normalization;
10. TMM —The trimmed mean of M values;
11. NCBI — National Center for Biotechnology Information.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Transcript profiling for early stages during embryo development in Scots pine / Irene Merino, Malin Abrahamsson, Lieven Sterck et al. // *BMC plant biology.* — 2016. — Vol. 16, no. 1. — P. 255.

[2] Comparative analysis of transcript abundance in Pinus sylvestris after challenge with a saprotrophic, pathogenic or mutualistic fungus / Aleksandra Adomas, Gregory Heller, Åke Olson et al. // *Tree physiology.* — 2008. — Vol. 28, no. 6. — Pp. 885–897.

[3] *De Heredia, Unai López.* RNA-seq analysis in forest tree species: bioinformatic problems and solutions / Unai López De Heredia, José Luis Vázquez-Poletti // *Tree Genetics & Genomes.* — 2016. — Vol. 12, no. 2. — P. 30.

[4] *Viola, Ivana L.* Structure and evolution of plant homeobox genes / Ivana L Viola, Daniel H Gonzalez // Plant Transcription Factors. — Elsevier, 2016. — Pp. 101–112.

[5] What Do We Know about Homeodomain–Leucine Zipper I Transcription Factors? Functional and Biotechnological Considerations / Pamela A Ribone, Matías Capella, Agustín L Arce, Raquel L Chan // Plant Transcription Factors. — Elsevier, 2016. — Pp. 343–356.

[6] The rice WUSCHEL-related homeobox genes are involved in reproductive organ development, hormone signaling and abiotic stress response / Saifeng Cheng, Yulan Huang, Ning Zhu, Yu Zhao // *Gene.* — 2014. — Vol. 549, no. 2. — Pp. 266–274.

[7] Analysis of the WUSCHEL-RELATED HOMEOBOX gene family in the conifer *Picea abiesreveals* extensive conservation as well as dynamic patterns / Harald Hedman, Tianqing Zhu, Sara von Arnold, Joel J Sohlberg // *BMC Plant Biology.* — 2013. — Vol. 13, no. 1. — P. 89.

[8] Comparative expression pattern analysis of WUSCHEL-related homeobox 2 (WOX2) and WOX8/9 in developing seeds and somatic embryos of

the gymnosperm *Picea abies* / Joakim Palovaara, Henrik Hallberg, Claudio Stasolla, Inger Hakman // *New Phytologist*. — 2010. — Vol. 188, no. 1. — Pp. 122–135.

[9] *Palovaara, Joakim*. Conifer WOX-related homeodomain transcription factors, developmental consideration and expression dynamic of WOX2 during *Picea abies* somatic embryogenesis / Joakim Palovaara, Inger Hakman // *Plant Molecular Biology*. — 2008. — Vol. 66, no. 5. — Pp. 533–549.

[10] *Guo, Mei*. Down-regulation of a homeodomain-leucine zipper i-class homeobox gene for improved plant performance. — 2018. — US Patent App. 15/599,830.

[11] Genome-wide analysis of homeobox gene family in legumes: identification, gene duplication and expression profiling / Annapurna Bhattacharjee, Rajesh Ghangal, Rohini Garg, Mukesh Jain // *PLoS One*. — 2015. — Vol. 10, no. 3.

[12] Homeodomain–leucine zipper transcription factors: structural features of these proteins, unique to plants / Matías Capella, Pamela A Ribone, Agustín L Arce, Raquel L Chan // Plant transcription factors. — Elsevier, 2016. — Pp. 113–126.

[13] *Gehring, Walter J.* Homeo boxes in the study of development / Walter J Gehring // *Science*. — 1987. — Vol. 236, no. 4806. — Pp. 1245–1252.

[14] *Ades, Sarah E.* Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex / Sarah E Ades, Robert T Sauer // *Biochemistry*. — 1995. — Vol. 34, no. 44. — Pp. 14601–14608.

[15] Homeodomain-DNA recognition / Walter J Gehring, Yan Qiu Qian, Martin Billeter et al. // *Cell*. — 1994. — Vol. 78, no. 2. — Pp. 211–223.

[16] *Wolberger, Cynthia*. Homeodomain interactions / Cynthia Wolberger // *Current opinion in structural biology*. — 1996. — Vol. 6, no. 1. — Pp. 62–68.

[17] *Moens, Cecilia B.* Hox cofactors in vertebrate development / Cecilia B Moens, Licia Selleri // *Developmental biology.* — 2006. — Vol. 291, no. 2. — Pp. 193–206.

[18] The developmental gene Knotted-1 is a member of a maize homeobox gene family / Erik Vollbrecht, Bruce Veit, Neelima Sinha, Sarah Hake // *Nature.* — 1991. — Vol. 350, no. 6315. — Pp. 241–243.

[19] The true story of the HD-Zip family / Federico D Ariel, Pablo A Manavella, Carlos A Dezar, Raquel L Chan // *Trends in plant science.* — 2007. — Vol. 12, no. 9. — Pp. 419–426.

[20] *Mukherjee, Krishanu.* A comprehensive classification and evolutionary analysis of plant homeobox genes / Krishanu Mukherjee, Luciano Brocchieri, Thomas R Bürglin // *Molecular biology and evolution.* — 2009. — Vol. 26, no. 12. — Pp. 2775–2794.

[21] *Jain, Mukesh.* Genome-wide identification, classification, evolutionary expansion and expression analyses of homeobox genes in rice / Mukesh Jain, Akhilesh K Tyagi, Jitendra P Khurana // *The FEBS journal.* — 2008. — Vol. 275, no. 11. — Pp. 2845–2861.

[22] Genome-wide identification, expansion, and evolution analysis of homeobox genes and their expression profiles during root development in carrot / Feng Que, Guang-Long Wang, Tong Li et al. // *Functional & integrative genomics.* — 2018. — Vol. 18, no. 6. — Pp. 685–700.

[23] Genome-wide identification, classification, and expression pattern of homeobox gene family in Brassica rapa under various stresses / Nadeem Khan, Chun-mei Hu, Waleed Amjad Khan et al. // *Scientific reports.* — 2018. — Vol. 8, no. 1. — Pp. 1–17.

[24] Gene family structure, expression and functional analysis of HD-Zip III genes in angiosperm and gymnosperm forest trees / Caroline L Côté, Francis Boileau, Vicky Roy et al. // *BMC Plant Biology.* — 2010. — Vol. 10, no. 1. — P. 273.

[25] KNOTTED1-like homeobox genes of a gymnosperm, Norway spruce, expressed during somatic embryogenesis / Helena I Hjortswang, Annika Sundås Larsson, Geeta Bharathan et al. // *Plant Physiology and Biochemistry.* — 2002. — Vol. 40, no. 10. — Pp. 837–843.

[26] *Ahuja, M Raj.* Evolution of genome size in conifers / M Raj Ahuja, David B Neale // *Silvae genetica.* — 2005. — Vol. 54, no. 1-6. — Pp. 126–137.

[27] Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation / Jill L Wegrzyn, John D Liechty, Kristian A Stevens et al. // *Genetics.* — 2014. — Vol. 196, no. 3. — Pp. 891–909.

[28] Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies / David B Neale, Jill L Wegrzyn, Kristian A Stevens et al. // *Genome biology.* — 2014. — Vol. 15, no. 3. — P. R59.

[29] *Haas, BJ.* De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with trinity. Nat Protocol 8 (8): 1494–1512. — 2013.

[30] Identification of differentially expressed genes in rna-seq data of arabidopsis thaliana: A compound distribution approach / Arfa Anjum, Seema Jaggi, Eldho Varghese et al. // *Journal of Computational Biology.* — 2016. — Vol. 23, no. 4. — Pp. 239–247.

[31] *Robinson, Mark D.* A scaling normalization method for differential expression analysis of RNA-seq data / Mark D Robinson, Alicia Oshlack // *Genome biology.* — 2010. — Vol. 11, no. 3. — P. R25.

[32] *Wilkinson, Leland.* The history of the cluster heat map / Leland Wilkinson, Michael Friendly // *The American Statistician.* — 2009. — Vol. 63, no. 2. — Pp. 179–184.

[33] Utilization of tissue ploidy level variation in de novo transcriptome assembly of Pinus sylvestris / Dario I Ojeda, Tiina M Mattila, Tom Ruttink

et al. // *G3: Genes, Genomes, Genetics.* — 2019. — Vol. 9, no. 10. — Pp. 3409–3421.

[34] *Andrews, Simon.* A quality control tool for high throughput sequence data. 2010 / Simon Andrews, A FastQC // *Google Scholar.* — 2015.

[35] *Bolger, Anthony M.* Trimmomatic: a flexible trimmer for Illumina sequence data / Anthony M Bolger, Marc Lohse, Bjoern Usadel // *Bioinformatics.* — 2014. — Vol. 30, no. 15. — Pp. 2114–2120.

[36] BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs / Felipe A Simão, Robert M Waterhouse, Panagiotis Ioannidis et al. // *Bioinformatics.* — 2015. — Vol. 31, no. 19. — Pp. 3210–3212.

[37] *Finn, Robert D.* HMMER web server: interactive sequence similarity searching / Robert D Finn, Jody Clements, Sean R Eddy // *Nucleic acids research.* — 2011. — Vol. 39, no. suppl_2. — Pp. W29–W37.

[38] *Eddy, Sean.* HMMER user's guide. biological sequence analysis using profile hidden Markov models / Sean Eddy. — 2003.

[39] The Pfam protein families database / Alex Bateman, Ewan Birney, Lorenzo Cerruti et al. // *Nucleic acids research.* — 2002. — Vol. 30, no. 1. — Pp. 276–280.

[40] *Robinson, Mark D.* edgeR: a Bioconductor package for differential expression analysis of digital gene expression data / Mark D Robinson, Davis J McCarthy, Gordon K Smyth // *Bioinformatics.* — 2010. — Vol. 26, no. 1. — Pp. 139–140.

[41] *Li, Bo.* RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome / Bo Li, Colin N Dewey // *BMC bioinformatics.* — 2011. — Vol. 12, no. 1. — P. 323.

[42] A new transcriptome and transcriptome profiling of adult and larval tissue in the box jellyfish Alatina alata: an emerging model for studying venom,

vision and sex / Cheryl Lewis Ames, Joseph F Ryan, Alexandra E Bely et al. // *BMC genomics*. — 2016. — Vol. 17, no. 1. — P. 650.

[43] Basic local alignment search tool / Stephen F Altschul, Warren Gish, Webb Miller et al. // *Journal of molecular biology*. — 1990. — Vol. 215, no. 3. — Pp. 403–410.

[44] *Consortium, UniProt*. Activities at the universal protein resource (UniProt) / UniProt Consortium // *Nucleic acids research*. — 2014. — Vol. 42, no. D1. — Pp. D191–D198.

[45] *Du Plessis, Louis*. The what, where, how and why of gene ontology—a primer for bioinformaticians / Louis Du Plessis, Nives Škunca, Christophe Dessimoz // *Briefings in bioinformatics*. — 2011. — Vol. 12, no. 6. — Pp. 723–735.

[46] *Elhiti, Mohamed*. Structure and function of homodomain-leucine zipper (HD-Zip) proteins / Mohamed Elhiti, Claudio Stasolla // *Plant signaling & behavior*. — 2009. — Vol. 4, no. 2. — Pp. 86–88.

[47] Arabidopsis HD-Zip II transcription factors control apical embryo development and meristem function / Luana Turchi, Monica Carabelli, Valentino Ruzza et al. // *Development*. — 2013. — Vol. 140, no. 10. — Pp. 2118–2129.

[48] Comprehensive analysis of WOX genes uncovers that WOX13 is involved in phytohormone-mediated fiber development in cotton / Peng He, Yuzhou Zhang, Hao Liu et al. // *BMC plant biology*. — 2019. — Vol. 19, no. 1. — P. 312.

[49] *Chew, William*. Role of homeodomain leucine zipper (HD-Zip) IV transcription factors in plant development and plant protection from deleterious environmental factors / William Chew, Maria Hrmova, Sergiy Lopato // *International journal of molecular sciences*. — 2013. — Vol. 14, no. 4. — Pp. 8122–8147.

[50] Expression dynamics of WOX genes mark cell fate decisions during early embryonic patterning in Arabidopsis thaliana / Achim Haecker, Rita Groß-Hardt, Bernd Geiges et al. // *Development*. — 2004. — Vol. 131, no. 3. — Pp. 657–668.

[51] Analysis of the WUSCHEL-RELATED HOMEOBOX gene family in Pinus pinaster: New insights into the gene family evolution / José M Alvarez, Natalia Bueno, Rafael A Cañas et al. // *Plant Physiology and Biochemistry*. — 2018. — Vol. 123. — Pp. 304–318.

[52] The WUSCHEL-RELATED HOMEOBOX 3 gene Pa WOX 3 regulates lateral organ formation in Norway spruce / José M Alvarez, Joel Sohlberg, Peter Engström et al. // *New Phytologist*. — 2015. — Vol. 208, no. 4. — Pp. 1078–1088.

[53] Conserved factors regulate signalling in Arabidopsis thaliana shoot and root stem cell organizers / Ananda K Sarkar, Marijn Luijten, Shunsuke Miyashima et al. // *Nature*. — 2007. — Vol. 446, no. 7137. — Pp. 811–814.

[54] Molecular evolution and gene expression differences within the HD-Zip transcription factor family of Zea mays L. / Hude Mao, Lijuan Yu, Zhanjie Li et al. // *Genetica*. — 2016. — Vol. 144, no. 2. — Pp. 243–257.

[55] The Arabidopsis Athb-2 and-4 genes are strongly induced by far-red-rich light / Monica Carabelli, Giovanna Sessa, Simona Baima et al. // *The Plant Journal*. — 1993. — Vol. 4, no. 3. — Pp. 469–479.

[56] Shade avoidance responses are mediated by the ATHB-2 HD-zip protein, a negative regulator of gene expression / Corinna Steindler, Antonella Matteucci, Giovanna Sessa et al. // *Development*. — 1999. — Vol. 126, no. 19. — Pp. 4235–4245.

[57] The HAT2 gene, a member of the HD-Zip gene family, isolated as an auxin inducible gene by DNA microarray screening, affects auxin response in

Arabidopsis / Shinichiro Sawa, Maki Ohgishi, Hideki Goda et al. // *The Plant Journal*. — 2002. — Vol. 32, no. 6. — Pp. 1011–1022.

[58] A dynamic balance between gene activation and repression regulates the shade avoidance response in Arabidopsis / Giovanna Sessa, Monica Carabelli, Massimiliano Sassi et al. // *Genes & development*. — 2005. — Vol. 19, no. 23. — Pp. 2811–2815.

[59] Characterization of the class IV homeodomain-leucine zipper gene family in Arabidopsis / Miyuki Nakamura, Hiroshi Katsumata, Mitsutomo Abe et al. // *Plant physiology*. — 2006. — Vol. 141, no. 4. — Pp. 1363–1375.

[60] Transcription Factors BLH2 and BLH4 Regulate Demethylesterification of Homogalacturonan in Seed Mucilage / Yan Xu, Yiping Wang, Xiaoyu Wang et al. // *Plant Physiology*. — 2020. — Vol. 183, no. 1. — Pp. 96–111.

[61] WUSCHEL-RELATED HOMEOBOX 4 (WOX 4)-like genes regulate cambial cell division activity and secondary growth in Populus trees / Melis Kucukoglu, Jeanette Nilsson, Bo Zheng et al. // *New Phytologist*. — 2017. — Vol. 215, no. 2. — Pp. 642–657.

[62] WOX4 promotes procambial development / Jiabing Ji, Josh Strable, Rena Shimizu et al. // *Plant physiology*. — 2010. — Vol. 152, no. 3. — Pp. 1346–1356.

[63] KNAT2/6b, a class I KNOX gene, impedes xylem differentiation by regulating NAC domain transcription factors in poplar. / Yanqiu Zhao, Xueqin Song, Houjun Zhou et al. // *The New phytologist*. — 2019.

[64] *Hay, Angela*. KNOX genes: versatile regulators of plant development and diversity / Angela Hay, Miltos Tsiantis // *Development*. — 2010. — Vol. 137, no. 19. — Pp. 3153–3165.

[65] Phyllotactic pattern and stem cell fate are determined by the Arabidopsis homeobox gene BELLRINGER / Mary E Byrne, Andrew T Groover,

Joseph R Fontana, Robert A Martienssen // *Development*. — 2003. — Vol. 130, no. 17. — Pp. 3941–3950.

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ:
заведующий кафедрой

_____ В. А. Кратасюк
"22" июня 2020 г.

БАКАЛАВРСКАЯ РАБОТА

03.03.02 Физика

ГОМЕОБОКС-СОДЕРЖАЩИЕ ГЕНЫ В ИЗУЧЕНИИ ПРОЦЕССА РАЗВИТИЯ СОСНЫ ОБЫКНОВЕННОЙ PINUS SYLVESTRIS

Руководитель: 19.06.2020, _____ д.ф.-м.н., проф. М. Г. Садовский
дата, подпись уч.степень, должность

Выпускник: 20.06.2020, _____ Т. А. Гусева
дата, подпись

Красноярск 2020